AMDA, Automated Multi-Dataset Analysis: A web-based service provided by the CDPP.

Jacquey[1] C., V. Génot[1], E. Budnik[2], R. Hitier[3], M. Bouchemit[1], M. Gangloff[1], A. Fedorov[1], B. Cecconi[4], N. André[1], B. Lavraud[1], C. Harvey[1], F. Dériot[5], D. Heulet[5], E. Pallier[1], E. Penou[1] and J.L. Pinçon[6].

[1]: CDPP/CESR, CNRS/Université Paul Sabatier, 9, avenue du colonel Roche, 31028 Toulouse, France.
[2]: NOVELTIS, 2, Avenue Europe, 31520 Ramonville Saint Agne, France
[3]: Co-Libri, Cremefer 11290 Montréal, France
[4]: LESIA, Observatoire de Paris-Meudon, 5, place Janssen, 92195 Meudon
[5]: CNES, Centre spatial de Toulouse, 18 avenue E. Belin, 31401 Toulouse
[6]: LPCE, Laboratoire de Physique et Chimie de l'Environnement, 45071 Orléans, France

Abstract: We present AMDA (Automated Mutli-Dataset Analysis), a new service recently opened at CDPP. AMDA is a web-based facility for on line analysis of space physics data coming from either its local database or distant ones. This tool allows the user to perform on line classical manipulations such as data visualization, parameter computation or data extraction. AMDA also offers innovative functionalities such as event search on the content of the data in either visual or automated way, and the generation, use and management of time-tables. These time-tables can be seen as a brick of up-coming virtual observatories in space physics, and could be used as an input to extract data from other databases, Cluster Active Archive in particular.

1. Introduction.

For in depth studies of plasma objects such as the Earth's magnetosphere or the solar wind, it is necessary to analyse multi-points and multi-instruments measurements. In practice, that means that researchers have to exploit together data coming from many sources which can initially be heterogeneous in their internal organisation, their description or their cod-

ing format. At the scale of the individual researcher, or even of a laboratory, this requires an important investment and consumes a large amount of time and energy. This became critical when the ISTP program and the CLUSTER mission were launched. Facing this challenge, several data centres emerged with the support of the research institutions and the space agencies, such as in USA (SPDF, Space Physics Data Facilities), in Japan (DARTS, Data Archives and Transmission System) and in Europe (CAA, Cluster Active Archive).

The CDPP (http://cdpp.cesr.fr/, Centre de Données de Physique des Plasmas) is the french national centre for space physics data. It was jointly created by CNES (Centre National d'Etudes Spatiales) and CNRS (Centre National de la Recherche Scientifique) in 1998. Recently, the CDPP has opened a new web-based service called AMDA (Automated Multi-Dataset Analysis, http://cdpp-amda.cesr.fr/). This service offers "classical" functionalities such as data extraction, merging or visualisation but also innovative ones such as user-edited automated search or computation on the content of the data as well as the creation and the use of time-tables. The aim of this paper is to present this service. We first describe what are the needs met in space physics data analysis. Then we present the functionalities of AMDA and illustrate them with a use case of the study of the cross-tail current dynamics.

2. Analysis of the needs for data exploitation in space physics.

As schematically depicted in Figure 1, case studies or statistical studies in space physics are generally performed as follows: (1) The first step is generally the search of cases of interest. This search can be performed through examination of quicklooks or using quantitative criteria on the content of the data. (2) When an event or a set of events has been identified, the second step consists of retrieving and extracting the numerical data. (3) Before analysis with the user preferred software, the data may need to be formatted depending on their original format. This step benefits of the more and more common use of standard formats as CDF, CEF or NetCDF. (4) Then, basic treatments generally need to be applied e.g. bad value filtering, gap interpolation or merging data at the same baseline, … (5) In this step, the user performs various manipulations on the data. Many of them are specific but commonly include visualisation, parameter computation and comparison to standard models such as magnetic field models.

It is only when these steps have been accomplished that the interesting work, i.e., the interpretation of the observations, starts. Because there are many datasets available, most of which coming from different origins, formatted in various ways, the work required before interpreting the observations often consumes large amount of time and energy from both researchers and software engineers. This has direct impact on the data exploitation, i.e. on the scientific return of the experimental investment.

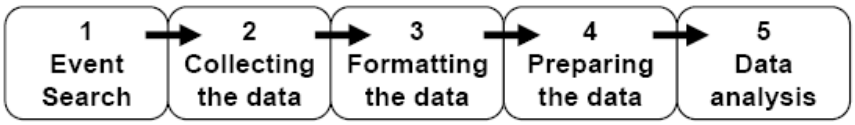

Figure 1: *Schematic illustration of the successive steps followed in a study based on data analysis in space physics.*

3. AMDA (Automated Multi-Dataset Analysis)

The main objective of AMDA is to help the user in performing the different steps described above. In order to enable complex operations on data, AMDA has been built around elementary objects: the *parameter* and the *time table*. A first and immediate difference relatively to most databases is that the concept of data files does not exist in AMDA at the user level. The user manipulates parameters without taking care of the files where they come from. Inside the AMDA system, the *parameter* is associated with properties (scalar, vector, tensor, units, …) and with corresponding options (decomposition, frame of reference, …). The second useful object in AMDA is the *time table* which is basically a collection of time intervals.

The functionalities of AMDA allow to use and to couple these two classes of objects. They are interactively activated through web-interfaces by the user who edits requests and collects the results in his/her workspace. All the requests and results may be saved for further sessions as soon as the user has registered. We now briefly describe these functionalities.

• **Parameter builder.** In the "My parameters" interface, the user can compute new parameters by editing mathematical expression combining existing parameters. Heterogeneous time bases are handled by AMDA transparently to the user. He/she freely chooses the time resolution of his/her

final new parameter and can also manage data gaps. Once a new parameter has been created it can be used as any pre-existing parameter.

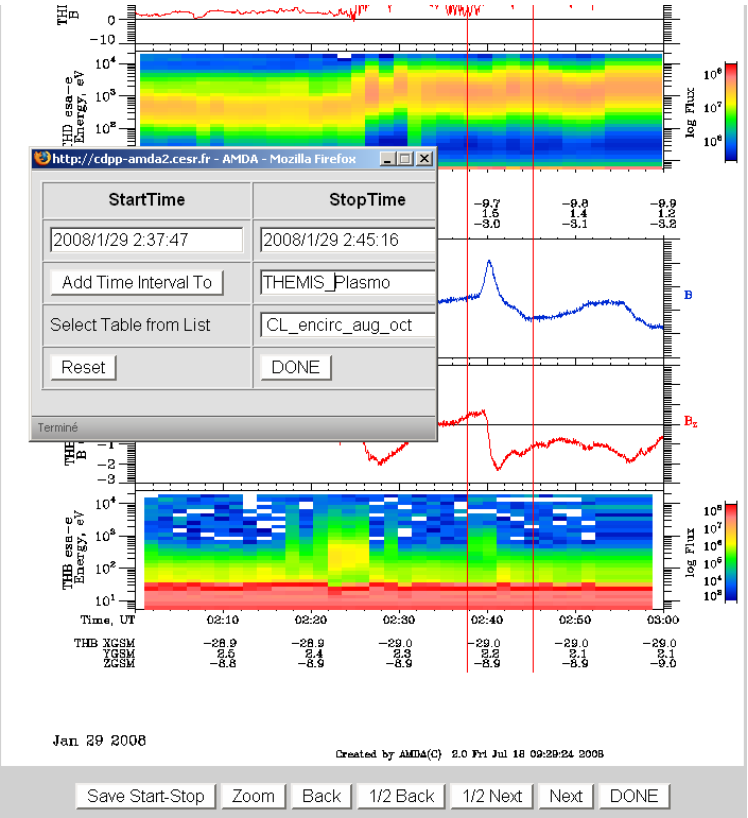

Figure 2. *Partial view of the "Plot Data" interface when the visual event selection functionality is activated. Magnetic field and electron spectra of THEMIS-B and THAMIS-D are represented here. A plasmoid event has been selected and recorded in the time table THEMIS_PLASMO.*

• **Remote data access.** Following the Virtual Observatory paradigm AMDA goes further in giving a direct access to parameters hold in distant databases (noticeably CDAWeb is available through AMDA). In the "External Data" interface, the user can browse through the parameters of the connected distant databases, select the desired ones and finally save them in his/her own *external data tree*. This data tree, with the usual hierarchy mission/instrument/dataset/parameter, will then be added to the existing local parameters. The parameters referenced in the external data tree can be used thereafter like any pre-existing ones.

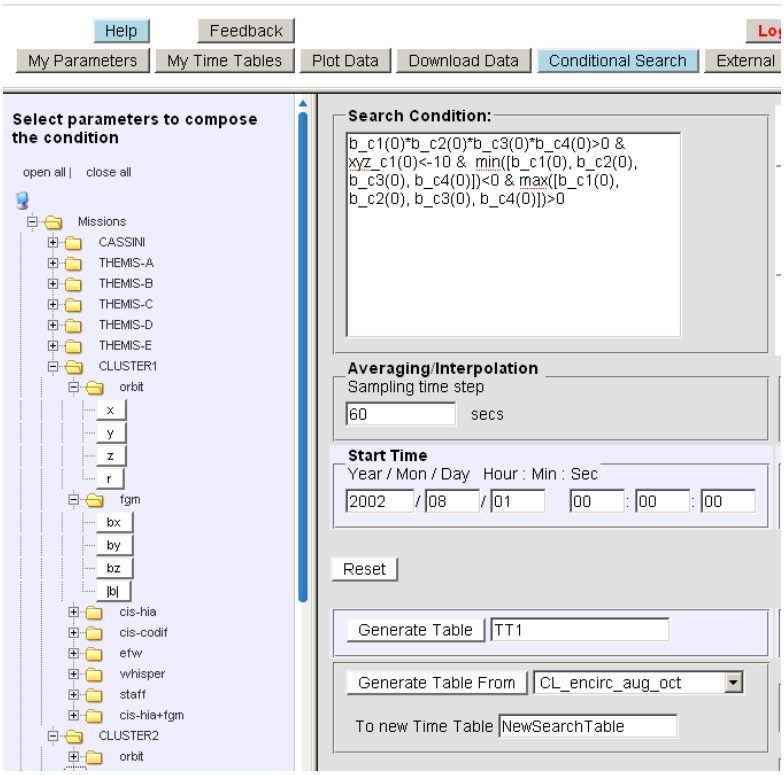

Figure 3. *Partial view of the "Conditional Search" interface. The criteria edited in the "Search Condition" window corresponds to the use case presented in section 4.*

• **Data download.** In the "Download Data" interface, newly created, remotely accessed or local parameters can be equally downloaded (or exported) by the user in different formats (plain ASCII, CEF, CDF) and with a chosen resolution. Time series may optionally be joined on a common time basis before exporting. The data merging is performed either by averaging or interpolating, depending on the time resolution of the original datasets. As it was mentioned above, the user does not need to care about new data downloads. It is fully automated according to the requested time interval and the parameters needed.

• **Data visualisation.** In the "Plot Data" interface, the user can edit a figure combining any available parameter with the desired options (reference frame, scaling,…). The request can be saved in such a way that the edited figure can be applied to any time interval chosen by the user. Note that the

interface provides the option to apply time shifting for solar wind propagation. Widgets also allow to shift the figure to the following/preceding time interval. This functionality also accepts time tables as input. In this latter case, widgets allow to skip the figure from a time interval to the next/previous one.

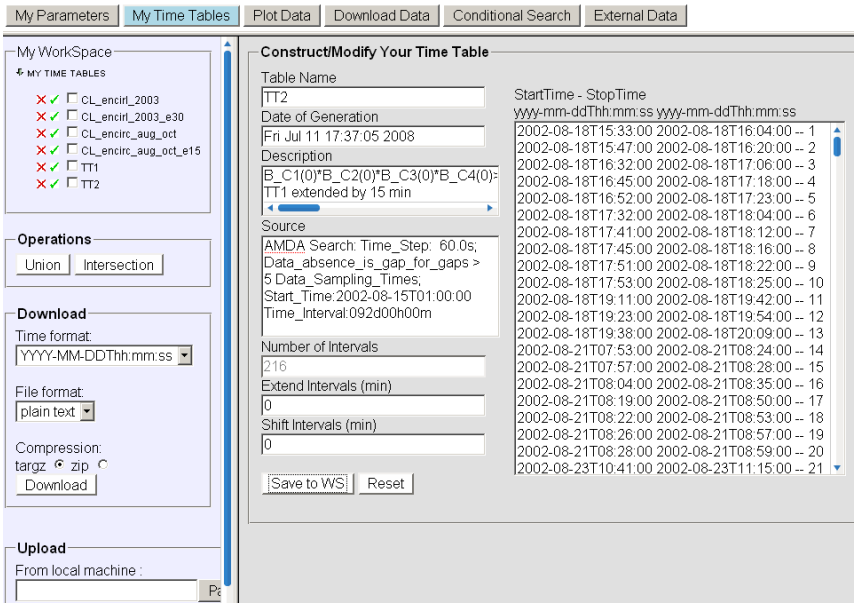

Figure 4. *Partial view of the "My Time Tables" interface.*

• **Visual search.** While browsing the requested figures, the user can retain intervals when a special feature, visually determined, occurs. He can do so by double clicking at the start/stop of the event (see Figure 2). Then an interface is provided to store this interval in a list; that is, the user can create a time table by visual inspection.

• **Automated conditional search.** A second way to create time tables is by selecting time intervals when a particular mathematical condition applied on given parameters is fulfilled. In the "Conditional Search" interface (see Figure 3), the user can edit his/her condition with mathematical functions and logical operands (>,<,&,|). This condition is applied on a given time window and only sub-intervals fulfilling the condition are retained to populate the time table. The time window may itself be a single time interval or a time table, offering the possibility to perform successive automated searches.

• **Time-table manager.** In the "My Time Tables" interface (see Figure 4), time tables may be subsequently edited and it is possible to apply operations on a single time table: extend/shrink/shift by a given duration; or on multiple time tables: union or intersection. The time table can be exported in ASCII format or in VOTables compliant to the IVOA standards.

4. A use case.

To illustrate the use of AMDA, let us consider the case of a user who wants to study from the CLUSTER data the local properties of the cross-tail current inside the neutral sheet with respect to interplanetary conditions and geomagnetic activity. He wants to focus on events when the current sheet is encircled by the CLUSTER satellites. As illustrated in Figure 5, a possible way to render this situation is to consider that this condition is fulfilled when two spacecraft are above the neutral sheet (i.e. BX >0) and the two other ones are below (i.e., BX<0).

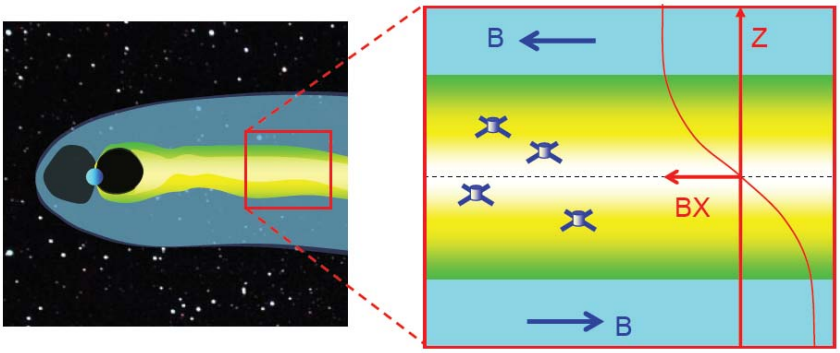

Figure 5. *Illustration of the neutral sheet encirclement by the CLUSTER constellation. The profile of the BX component through the current sheet is shown on the right side of the right panel. BX is positive above the neutral sheet and negative below.*

In the "Conditional Search" interface of AMDA (Figure 2), the user first performs an automated search of the time intervals when the CLUSTER spacecraft encircled the neutral sheet as defined above. He/she edits the following criteria:

$$B_{X1} \cdot B_{X2} \cdot B_{X3} \cdot B_{X4} > 0 \;\; AND \; \min\{B_{X1}, B_{X2}, B_{X3}, B_{X4}\} < 0$$
$$AND \; \max\{B_{X1}, B_{X2}, B_{X3}, B_{X4}\} > 0 \; AND \; X_{GSM1} < -10$$

He/she applies them to the CLUSTER data during the tail seasons, for example the one of the year 2002 (August to October). To avoid selecting too brief events, the filter is applied by computing on the fly 1 minute-averaged FGM data. AMDA then generates a first time table (hereafter TT1) listing all the time-periods when this neutral sheet encirclement condition is fulfilled. In the case of the 2002 tail season, 216 events are found.

Starting from this time table TT1, the user can then undertake several possible actions. Some examples:

• **Extraction of sub-database corresponding to CLUSTER neutral sheet encirclement events.** In the 'My parameter" interface, the user can formulate additional parameters such as the plasma beta, the total pressure for example. Then, in the "Download Data" interface, he/she can select the parameters of interest for his/her study and then download them in merged files corresponding to the time-intervals recorded in the time-table TT1. So, he/she constitutes a sub-database of the CLUSTER neutral sheet encirclement events that he/she can statistically treat off-line with his/her own tools and software. In view of analysing the context of these events, the user can generate a new time table TT2 by extending the time intervals of TT1 and download the corresponding complementary sub-database.

• **Extraction of corresponding sub-database of solar wind data.** To download the interplanetary data obtained by ACE, the user can first extend the Time-Table by 60 minutes and shift it by 45 minutes, corresponding to a rough averaged time-delay of solar wind propagation. (The on-the-fly computation of the solar wind propagation time-delay is an up-coming functionality of AMDA).

• **Filtering the event collection with additional criteria.** As explained in section 3, the "Conditional Search" interface allows to perform successive searches. The user can use this possibility if for instance he/she wants to:

(i) find the events when all the CLUSTER spacecraft were embedded in the central plasma sheet (with a condition on the plasma beta parameter),

(ii) classify the events as a function of their location in the XY_{GSM} plane, the observed flows, the wave activity level, the geomagnetic activity…

(ii) find the events when GEOTAIL and/or POLAR were also imbedded in the plasma sheet (with condition on their location and their plasma measurement) in order to perform multi-scale studies.

These automated searches would generate new time tables which could be thereafter used to constitute new sub-databases or to produce figures via the "Plot Data" interface.

5. Conclusion and perspectives.

AMDA is a web-based service developed with the aim to help researchers in space physics. AMDA offers functionalities to access and analyse multi-dataset in a transparent fashion and allows perform event search. This tool has already been used in various studies (Louarn et al., 2006, Génot et al., 2008,a,b). AMDA is in continuous development. Future developments concern new analysis tools, new products, enhancement of interoperability and the creation/management of catalogues.

AMDA is evolving in the Virtual Observatory paradigm. It gives a direct access to data from distant databases and includes a connection layer compliant with the SPASE (http://www.spase-group.org/) standards. AMDA produces, manipulates and uses time tables. The time tables can be seen as one of the primary brick to be used for the interoperable exchanges in space physics. This idea encouraged to start a collaboration between databases (AMDA and CAA) on one hand, and generic tools (QSAS and CL) on the other hand.

Acknowledgements

The authors would like to thank P. Louarn, C. Perry, T. Allen, S.F. Fung and D. Bilitza for helpful discussion and the CDPP User Committee for having tested AMDA.

References

Génot V., E. Budnik, C. Jacquey, I. Dandouras, E. Lucek,, (2008), Mirror mode events observed with Cluster in the Earth magnetosheath : statistical study and IMF/solar wind dependence, Advances in Geosciences, in press

Génot V., E. Budnik, P. Hellinger, T. Passot, G. Belmont, P. Travnicek, P.-L. Sulem, E. Lucek, and I. Dandouras, Mirror structures above and below the linear instability threshold : Cluster observations, fluid model and hybrid simulations, submitted to Annales Geophysicae

Louarn P., C. Jacquey, E. Budnik, and V. Genot, (2006) the CDPP, FGM, CIS and STAFF teams,The active plasma sheet: definition of 'events' and statistical analysis , Proceedings of the 8th International Conference on Substorms, March 27-31,2006, Banff Centre, *Editors M. Syrjaeso and E. Donovan*